*Missouri Secretary of State Robin Carnahan*
*Records Services Division*
*presents:*

## Electronic Records: Preservation and Access

**Workshop 7 in the Missouri Electronic Records Education and Training Initiative**

**October 6, 2005**

**Presented by:**
**Charles M. Dollar**

Provided under contract with:

***e*Visory**

## Workshop Overview

1. What is digital archiving?
2. Why is digital archiving so difficult?
3. Digital archiving standards
4. Digital arching and technology obsolescence
5. Digital archiving special issues
6. Operational digital archiving programs
7. Digital archiving programs to watch
8. A digital archiving microfilm alternative
9. Putting everything together

2

Peter Lyman:

**"Are digital signals destined to be a kind of oral culture, living only as long as they are remembered and repeated? What can be done to preserve our digital cultural heritage?"** *Time & Bits: Managing Data Continuity* (1998)

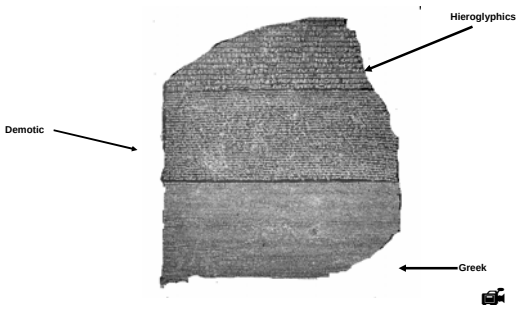
3

## Section 1

## What Is Digital Archiving?

4

**Rosetta Stone (196 BC)**




5

## Digital Object Attributes

- Physical
- Logical
- Conceptual

6

## Physical Attributes of Digital Records

- Signals on a storage medium
- Binary representation
- No inherent intelligence or meaning – a string of 1s and 0s

7

## What Is This?

01110111011110000111100100110001
00110010001100110011010000110101
01110000001100100100000100110101
00110011001101010011000100110101
00110100001101110011100000111001
00111000001101110011000000110000
00110010001100110011010000110100

8

## Logical Attributes

- Software applications recognize a logical object based on data type
- ASCII, native word processing, PDF, TIFF, JPEG

9

## A Clinical Trial Data Record

01110111011110000111100100110001
00110010001100110011010000110101
01110000001100100100000100110101
00110011001101010011000100110101
00110100001101110011100000111001
00111000001101110011000000110000
00110010001100110011010000110100

XYZ12345P2A535I54789870023440002156799

10

## Conceptual Attributes

- Conceptual objects are real
- Rendering in a human understandable form
- Content, structure, context
- Multiple ways of rendering
  - Word 2000 PDF
  - Tiff images & JPEG images
  - HTML & XHTML

11

## A Clinical Trial Data Record

**Structure & Context Metadata**

XYZ12345P2A535I54789870023440002156799

| | Field Definitions (digital) |
|---|---|
| Project No. | XYZ12345 |
| Protocol | P2A5351 |
| Investigator | 47897700 |
| Patient ID | 23440 |
| Visit No. | 002 |
| Result | 1576799 |

12

## A Clinical Trial Data Record

**Structure & Context Metadata**

XYZ12345P2A535I5478987002344000 2156799

| | Field Definitions (digital) | Code Rules Definitions |
|---|---|---|
| Project No. | XYX12345 | Research |
| Protocol | P2A5351 | Happy Daze |
| Investigator | 47897700 | Dr. Nuzzle |
| Patient ID | 23440 | |
| Visit No. | 002 | Cumulative |
| Result | 1576799 | Euphoric |

13

## A Clinical Trial Data Record

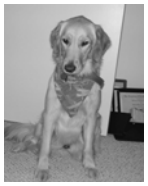
**Structure & Context Metadata**

XYZ12345P2A535I5478987002344000 2156799

| | Field Definitions (digital) | Code Rules Definitions | Derived Values (digital) |
|---|---|---|---|
| Project No. | XYX12345 | Research | |
| Protocol | P2A5351 | Happy Daze | |
| Investigator | 47897700 | Dr. Nuzzle | |
| Patient ID | 23440 | | |
| Visit No. | 002 | Cumulative | |
| Result | 1576799 | Euphoric | |

14

## Sophie



15

## Evolution of Digital Archiving

- 1970's focus on preservation of storage media
- 1986 redefinition of preservation
- 1991 "continuing processibility"
- 1992 "access as far into the future as necessary" (Pat Battin)
- 1999 "ensuring access to processible authentic (trustworthy) electronic records for as far into the future as necessary

16

## Digital Archiving

- Protect records from
  - Loss
  - Alteration
  - Corruption
- Ensure long-term accessibility
  - Across organizational boundaries
  - Across multiple technology changes and environments
- Allow future users
  - Retrieve, access, view, etc
  - Data, documents, records, etc
  - Multiple ways and purposes
  - Retain meaning and authenticity
- Recording medium (e.g., CD-ROM)
- Operating system (e.g., Windows)
- Storage specification (e.g., ASCII, sound, images)
- Data coding system (e.g., ASCII, XML, PDF)
- Metadata (context, stylistic)
- All of these change over time

17

## Processible Authentic Electronic Records

- Processible = reprocessibility
  - Replicate process
  - Spread sheets and databases
  - Snapshots and viewer technologies are inadequate
- Authentic
  - Trustworthy evidence
  - "Look and feel"
  - Quality assurance indicators

18

## Section 2

## Why Is Digital Archiving So Difficult?

19

## Digital Archiving Is Difficult

1. Digital v. traditional information artifacts
2. Digital records must be stored and retrieved digitally
3. Digital records require an "interpreter"
4. Hardware/software dependence
5. Technology obsolescence
6. Alternatives

20

## Digital v. Traditional Information Artifacts

- Traditional
  - Physical representation
  - Visible and self-contained
  - Require only final, human interpretation
- Digital
  - Logical representation
  - Invisible bit representation
  - Bits must be interpreted
  - Final human interpretation

21

## Digital Records Must Be Stored and Retrieved Digitally

- Recording medium (e.g., CD-ROM)
- Operating system (e.g., Windows)
- Software application (e.g., Microsoft Word)
- Storage specification (e.g., ASCII, sound, images)
- Data coding system (e.g., ASCII, XML, PDF)
- Metadata (context, stylistic, header data)

22

## Digital Records Require An "Interpreter"

- Bit streams are *not* self-explanatory
- Digital documents can only be interpreted by its software
- Software is necessary to see exactly what the author saw or did not see

23

## Software-dependent Documents Are Really System-dependent

**Journey of a Byte**

### Creation/Storage

- Key "A"
- Converted to ASCII Decimal 65 (01000001)
- Saved as Doc or TXT
- Operating system notifies drive controller
- Drive controller receive 01000001
- Recodes to 1100011100
- Writes to storage media

### Retrieval/Display

- Application sends request to disk controller
- Drive controller notifies drive to find "A" and copy it
- Drive strips off recoded values to original 65 (01000001)
- Notifies Operating System to pick up the "A"
- Operating system delivers 01000001 to the App. System
- App. recognizes ASCII text format
- Renders "A" for viewing

24

## Technology Obsolescence

- Superseded or displaced technical solutions
- Inevitable and irreversible
- Induced technology obsolescence
- Real technology obsolescence

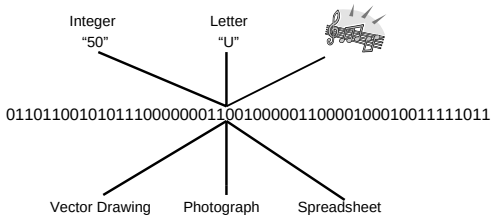
25

## Alternatives

- Retain software originally used to create documents
- Run similar software
- Translate document into a new file specification

26

## What Do Binary 1s and 0s (Representations) Represent?




27

## ASCII Digital Representation

| | | |
|---|---|---|
| C | 67 | 01000011 |
| h | 104 | 01110000 |
| a | 97 | 01100001 |
| r | 114 | 01110010 |
| l | 108 | 01101101 |
| e | 101 | 01100101 |
| s | 115 | 01110011 |

28

## Raw O-ring Data From Space Orbiter Launches, 1981 - 1985

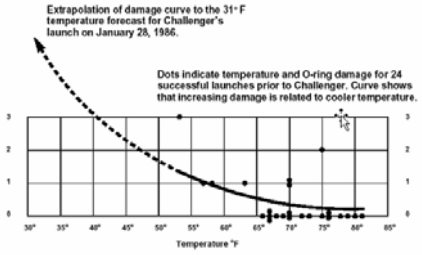
| Levels of O-ring damage | 53 | 57 | 58 | 63 | 66 | 67 | 68 | 69 | 70 | 72 | 73 | 75 | 76 | 78 | 79 | 80 | 81 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 | 1 | | | | | | | | | | | | | | | | |
| 2 | | | | | | | | | | | | 1 | | | | | |
| 1 | | 1 | 1 | 1 | | | | | 2 | | | | | | | | |
| 0 | | | | | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 1 | 1 | 1 | 1 |

Temperature °F

Source: Jeff Rothenberg

29

## Extrapolation of O-ring Damage, Orbiter Launches, 1981 - 1985




Source: Jeff Rothenberg

30

## Section 3

## Digital Archiving Standards

## Overview

- **ISO 9000** (1992) Quality Systems
- **ISO 15489** (2001) Information and documentation – Records management – Part 1 and 2
- **ISO 14721** (2003) *Space data and information transfer systems -- Open archival information system -- Reference model*
- **ISO/TR 14982** *(2005) Document Management Applications – Long-term preservation of electronic document-based information*

32

## ISO 9000

- Quality records (4.16)
  - Protect documentary evidence from being irretrievable
- Establish and maintain procedures for the maintenance of quality records
- Ensure that quality records are legible
- Store quality records in such a way as to prevent damage, deterioration, and loss

33

## ISO 15489

- Records principles
- Digital Storage
- Continuing retention

34

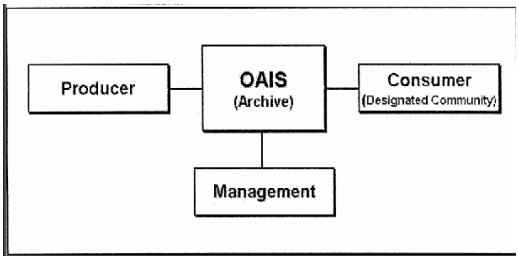
## ISO 14721 Open Archives Information System

Background

- Consultative Committee for Space Data Systems (1982)
- Formal standards for the long-term storage of digital data
- Reference model (1995-96)
- Open archival information system (OAIS)

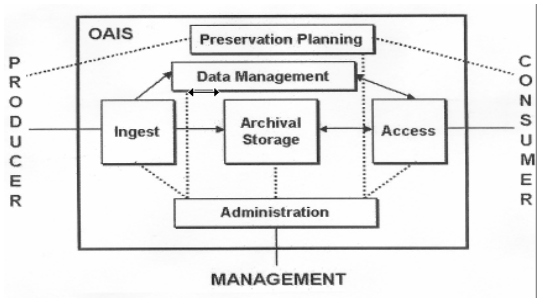
35

## OAIS Environment




36

## OAIS Functional Model




37

## Ingest

- Validation that the information is uncorrupted and complete
- Transformation into a form suitable for storage
- Extraction/creation of metadata
- Transfer to Archival Storage

38

## Archival Storage

- Ensure that the information resides in the appropriate form of storage
- Ensure that bit streams remain complete and reliable
- Media renewal
- Error checking
- Disaster recovery
- Receives access requests

39

## Data Management

- Maintains descriptive metadata
- Manages administrative data supporting internal OAIS operations
- Update databases as necessary
- Supports search/retrieval

40

## Preservation Planning

- Mapping out preservation strategy
- Monitors technology developments
- Monitors changes to accommodate expectations of Designated Community
- Recommends updating of policies and procedures as appropriate

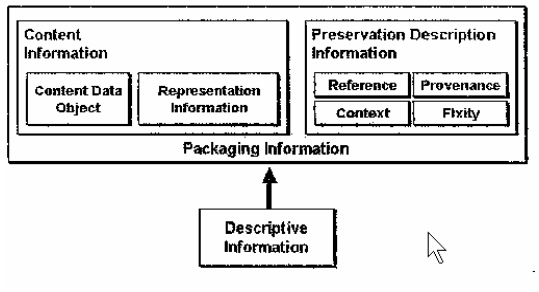
41

## Access

- Manages processes and services whereby users (Designated Community) locate, request, and receive “archived content”
- Interface with Designated Community

42

## Archival Information Package




43

## Archival Information Package Components

- Content data object can be any self-contained digital file
- Representation information
- Context information
- Provenance information
- Fixity

44

## Importance of OAIS Model

- A shared view of the core functional and informational requirements of digital archiving is essential
- A shared view facilitates the development of an interoperable network of digital archives
- An interoperable network of digital archives will be a key component in grid computing

45

## ISO/TR 18492

- TC 171 Document Management Applications
  - June 2005
  - High level steps and procedures
- *Authentic Electronic Records: Long-Term Access Strategies* (1999, 2000, 2002, and 2005)
- Review main elements

46

## Scope

- "Document-based information"
- Set aside as evidence for future use
- Narrow preservation focus
- Technology neutral information standards

47

## 18492 Definitions

- "Document-based information"
- "electronic archiving" (i.e. digital archiving)
- Long-term preservation
- Migration

48

## Goals of A Long-Term Preservation Strategy

- Readable
- Intelligible
- Identifiable
- Retrievable
- Understandable
- Authentic

49

## Authenticity Protection Approaches

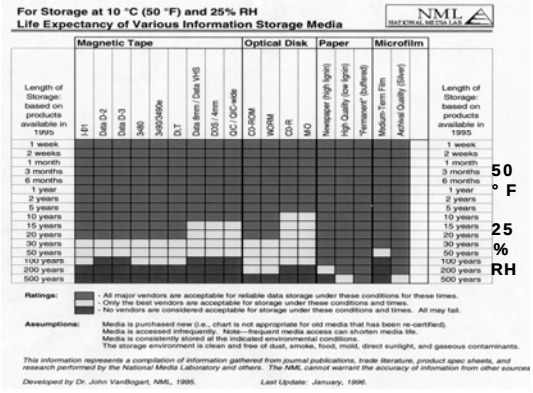
- Transfer from an operating environment to a trusted third party
- Storage environment
- Security protection

50

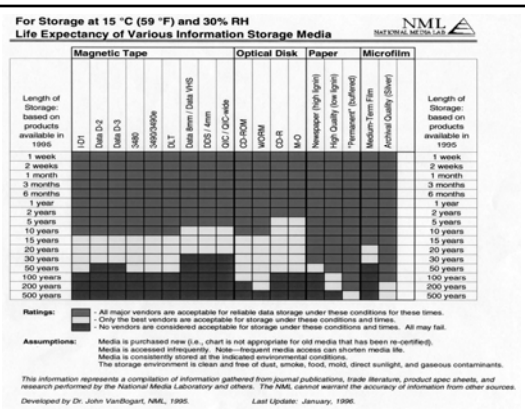
## An Audit Checklist for the Certification of Trusted Digital Repositories

- RLG/NARA – September 2005
- Repository functions
- Designated community and the usability of information
- Technology and technical infrastructure

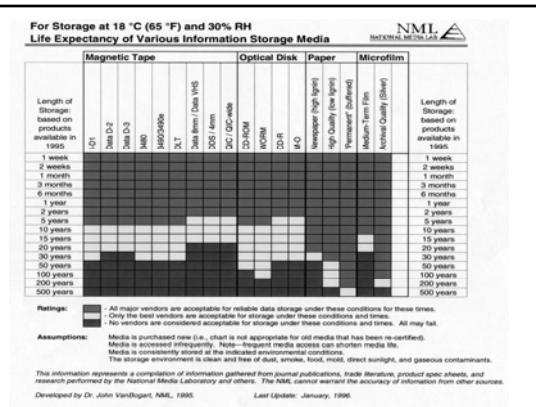
51



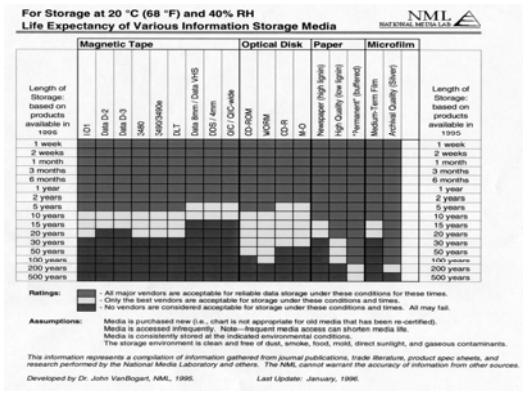

52



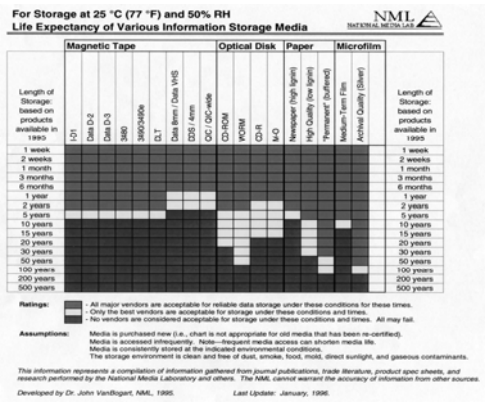

53




54




55




56

## Section 4

## Digital Archiving And Technology Obsolescence

## Overview

**Active Records**

- **Media Renewal**
- **Conversion**

**Legacy Records**

- **Media Renewal**
- **Digital Museum**
- **Emulation**
- **Migration**

58

## Maintaining Processibility

- Media renewal
  - Reformat
  - Copy
- Conversion

59

## Reformat

- A change in the underlying bit stream that occurs
  - physical carrier is changed
  - character code is transformed
  - fixed length to variable length records
- No change in the structure, content, or context

60

## Reformat - continued

- Access to application software is not necessary
- Not necessary for the software to know if the records are ASCII, Bitmap Images
- Examples
  - 6250 bpi tape to 3480 tape
  - EBCDIC to ASCII

61

## When To Reformat

- Time of transfer to “trusted third party”
- Storage devices and media upgrades
- Pre-established point in life expectancy of storage media

62

## When To Copy

- Time of transfer to a “trusted third party”
- Replacement of existing storage media
- Annual sample checked for “readability problems”

63

## Copy

- Exact duplication of the underlying bit stream (1s and 0s) that perpetuates the original representation, intellectual content, and context of electronic records
- Access to creating software application is not necessary
- Not necessary to know if the records are ASCII, Bitmap Images, Vector drawings, etc

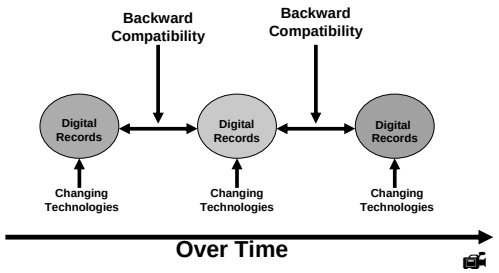
64

## Maintaining Processibility Through Conversion

- Automatic transfer of electronic records from one technology environment to another with no *undocumented* loss in structure and no loss in content or context
- “Automatic” means little or no user involvement is required. Export/import functionality recognizes the source or target software and carries out a reliable conversion (HTML to XHTML)

65

## Conversion Life Cycle Process




66

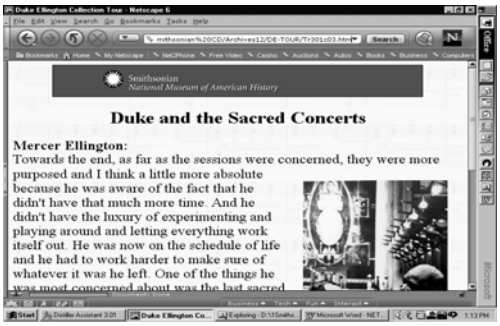
## Example of Active Records Conversion

World-Wide Active Duty Military Deaths

- Began as Combat Area Casualty File (Vietnam)
  - Korean War
  - All military activities to the present
- Conversions
  - 1950s punch cards
  - 1960s magnetic tape
  - 1980-81 integration of service into a HP relational database system
  - 1990-91 converted to an Access database system
  - Late 1990s converted to Oracle database system

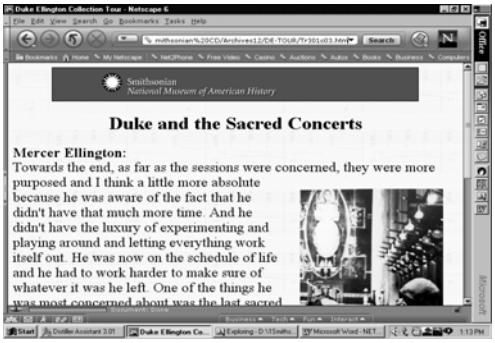
67

## HTML Page




68

## XHTML Page




69

## Other Examples of Conversion

- Rich Text Format (RTF)
- Word Perfect to Word
- Word 97 to Word 2003
- Word to PDF
- PDF to PDF

70

## Data Archaeology

- Seamus Ross and University of Glasgow
- Minimalist strategy
  - Ensure the continuing readability of electronic records
  - Collect documentation about software - operating systems and applications
  - Convert to new technology on demand (emulation on demand/reverse engineering)
  - Doom's Day Book

71

## Original Hardware & Software

- Museum Perspective
  - Retain an operational copy of software, operating system, and hardware
  - Jet Propulsion Lab
  - Washington Digital Archives
- Problems
  - Temporary solution at best
  - Today's v. tomorrow's technologies

72

## Viewer Technology

- View, copy, or print electronic records in various formats
- Does not replicate full functionality – Acrobat Reader, Quicken Plus, SnagIt
- Works best when evidential and information requirements can be satisfied by viewing a record
- Can be useful with legacy records
- Technology cul de sac

73

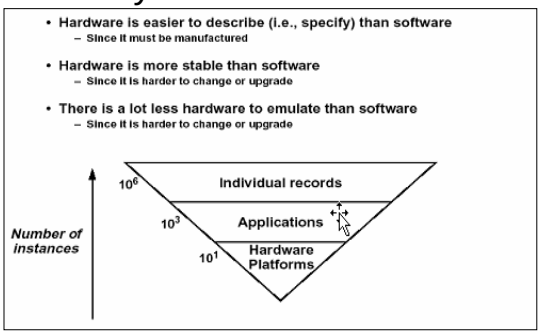
## Emulation

- Jeff Rothenberg, Rand Corporation
  "Emulation is a process in which one computer is used to reproduce the behaviour of another computer with such fidelity that the emulation can be used in place of the original computer"
- Supports executable "digital originals"
- Requires native application and emulator of original computer platform
- Presumes on-going readability of electronic records

74

## Why Emulate Hardware

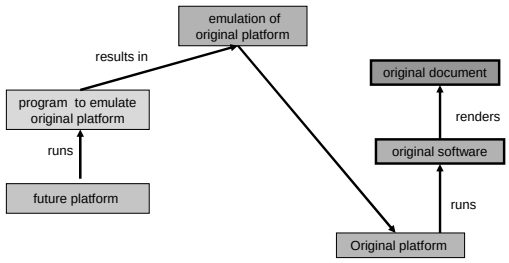
- **Hardware is easier to describe (i.e., specify) than software**
  - **Since it must be manufactured**
- **Hardware is more stable than software**
  - **Since it is harder to change or upgrade**
- **There is a lot less hardware to emulate than software**
  - **Since it is harder to change or upgrade**




Source: Jeff Rothenberg

75

## Overview of Emulation




76

## A Variation on Emulation

- Raymond Lorie, IBM Research Center
- Concepts
  - Virtual Machine
  - Universal Virtual Computer
  - Rendering of original document
- Demonstration at Netherlands Library
  - Extract data and structure from PDF files
  - Retain a bit map image
  - Reconstruct original file without PDF
    - Scroll
    - Word searches
    - Navigate document

77

## Program P

- Solution to obsolescence: write the program for a Universal Virtual Computer (UVC).
- The only thing needed in the future is an emulator of the virtual machine.
- UVC is a complete but simple computer.
- Simplicity will ensure longevity and facilitate the writing of a UVC emulator.

78

## UVC activities

- A UVC design exists, with a corresponding assembler and emulator.
- Proof of concept studies
  - (KB) Dutch National Library, for JPEG and PDF
  - ICTU (Branch of Dutch Gov't) spread sheets
- Currently looking at a wider set of formats, trying to generalize data extraction and encoding across formats

79

## UVC: Overall mechanism

Today:

- Store information in any format, together with a program P (the decoder).
- P produces a logical view (organized as an *XML-like* structure.

In the future:

- Run program P
- Use the logical view as input to new viewers and/or new application programs.

80

## Emulation Problems

- Users will have to know how to run obsolete software
- Likely to require vernacular copies
- Bit streams must not change
- May have to emulate more than processors
- No specifics yet on a commercially viable emulator specification/development method

81

## CAMiLEON Project: An Empirical Study

- Test the technical feasibility of using emulation to preserve digital information
- Evaluate the effectiveness of digital preservation strategies from the user's perspective
- University of Leeds (UK) Team – technical approaches to designing and implementing digital preservation environments
- University of Michigan (US) Team – unpacking concept of "significant properties" through modeling and user studies

Source: Cal Lee

82

## Scope of CAMiLEON

- Chuckie Egg
  - Video game
  - Popular in the UK
  - 1980s
- Complex Text Documents
  - Papers of James J. Duderstadt
  - 1990 – 1993
  - MORE Outliner on MacIntosh

Source: Cal Lee

83

## Chuckie Egg - Methodology

- 30 participants
- First spent one hour using original (old hardware and software) version
- Then switched to modern computer and used either emulated or migrated/transformed version

Source: Cal Lee

85

## Chuckie Egg - Findings

- Minor differences between the original and the emulated or migrated/transformed versions, including several not expected
- Observed differences related to the hardware environment, especially keyboard, sound, screen display, and interaction speed
- No statistically significant evidence that emulated version did a better job of preserving look and feel
- Participants generally preferred playing migrated/transformed or emulated version rather than original game on BBC Micro

Source: Cal Lee

86

## MORE Format

- One of several outliner programs available in 1990s
- Advanced functionality for the time – text wrapping, folding, user-defined rules, styles and comments, inheritance of properties
- Provided three different views of a document: outline, bulleted list, tree chart
- Supported until 1997
- Duderstadt used MORE extensively as an "idea processor"

Source: Cal Lee

87

## Duderstadt Papers – Methodology

- 30 participants
- Which documents did Duderstadt create?
  - Comparison of three speeches in one format
  - Comparison of same speech in three different formats
  - differences between the original MORE files and either Word or Text

Source: Cal Lee

88

## Duderstadt Papers - Findings

- 3 speeches in same format
  - insufficient information to determine the original
- Same speech in 3 formats
  - identified the original more accurately
  - better at determining how closely a version resembled the original
  - were more confident in their rankings
- MORE v. text/Word
  - greater differences between MORE and text
    v.
  - MORE and Word
- Original version
  - a major concern for only a small minority
  - Usability of the format being valued more

Source: Cal Lee

89

## Duderstadt Papers – Further Findings

- A majority (16 out of 30) incorrectly identified the original documents
- Speculations about capabilities of computers at the time were often incorrect
- Only one person used any of the metadata associated with documents to determine original

Source: Cal Lee

90

## Significance and Implications

- The "emulation vs. migration" dichotomy is not precise enough to make categorical digital preservation decisions
- Importance of contextual cues
  - original creation/use
  - purpose/audience
  - computing environment

Source: Cal Lee

91

## Migration

- Ensure usable and trustworthy electronic records for as long as necessary without regard for computer technology platform
- Presumes "readable" electronic records
- Involves converting electronic records to technology neutral file formats
- Requires backward compatibility
- Preserves processibility
- Potential risk

92

## Legacy Records Migration Overview

- What constitutes legacy records
  - Software/operating system dependent
  - No export functionality
  - Mapping requires extensive coding
- Brodie & Stonebreaker, *Migrating Legacy Systems* (1995)
  - Ten steps
  - Labor intensive
  - Costly
  - Possibility of failure

93

## Migration Technology

- Analyze legacy information system
- Decompose the legacy system structure
- Design the target interfaces
- Design the target applications
- Design the target database
- Install and test target environment
- Create necessary gateways
- Migrate the legacy records
- Migrate the legacy application
- Migrate the legacy interfaces

94

## NIPS File

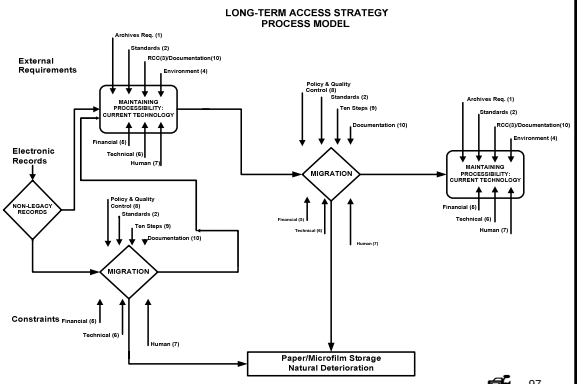
- Vietnam Military Records
- National Information Processing System
  - Report generating system
  - IBM designed for DOD
- Records
  - Each record contained field location data
  - Variable length
- Required use of NIPS software

95

## DeNIPSed Files

- Stripped off control characters (non-printing data)
- Created a flat file
  - Each table row represented a record
  - Each table column represented a data element
- All records were fixed length

96



97

## NIPS Combat Activities File

- Air Force bombing missions
  - Ho Chi Minh Trail, 1965 – 1975
  - 2,000,000 tons of bombs
  - At least 30% did not explode
- International effort to locate and defuse bombs
- NIPS Files
  - Bombing sorties
  - Map coordinates
  - DOD project (Roy Stanley)
  - Software to migrate data
  - Missing codebooks and values
  - "Data anomalies", 1976-78

98

## Legacy Migration Lessons Learned

- Fully understand legacy system and target system
- Flexible evolutionary approach is required
- Not possible in all cases to replicate exactly all data values
- Large volumes of data will be produced
- Expensive

99

## The Pros and Cons of Migration

Pro

- Methodology is understood and demonstrated – music transcription
- Utilizes today's technologies
- Builds on "backward compatibility" where possible
- Employs open standards
- Adaptable to many digital records

Con

- Standards are not a panacea
- Digital record bit stream changes
- Potential to introduce errors without quality control
- Complex inactive digital records migration is costly
- May lose some of original look and feel
- Never ending

100

## Section 5

## Digital Archiving Special Issues

## Overview

- Storage media
- File Formats
- Metadata

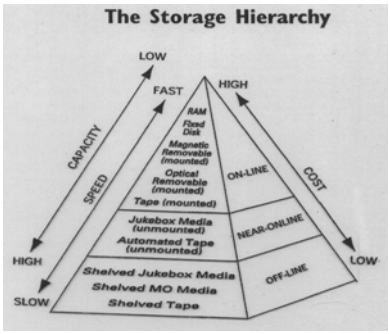
102

## Digital Information Storage Media

- Fixed magnetic and removable magnetic and optical media
- Tremendous growth in storage capacity
  - Megabytes
  - Gigabytes
  - Terabytes
  - Petabytes
- Corresponding decrease in cost

103

## Strategies for Selecting Digital Storage Media




104

## Digital Magnetic Media

- Direct access
  - Hard disk
  - Speedy access
    - available sectors and tracks
    - physical proximity is not required
- Sequential access
  - Magnetic tape
  - One record follows another
  - Relatively "slow" access time

105

## Fixed Rewriteable Magnetic Media

- Access time is measured in milliseconds
- Storage capacity ranges from 80GB to 1 TB
- Transfer rate ranges from 50 MB to 100 MB per second
- Cost is less than $1 per GB

106

## Spinning Disks

- Emerging consensus that "spinning disk" technology is more cost effective than magnetic tape. National Academy of Science, Digital Preservation Committee (2003)
- Several National Archives advocate "spinning disk" technology

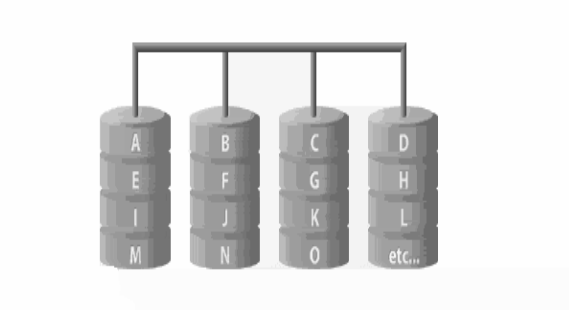
107

## What Is "Spinning Disk" Technology

- Non-removable hard disk?
- Continuously spinning?
- RAID Technology?
  - Spread data across multiple hard disks
  - At least five "flavors" called levels
    - Level 0 – Stripe blocks of data across two drives
    - Level 1 – Mirroring and Duplexing
    - Level 5 – Combines striping with parity check

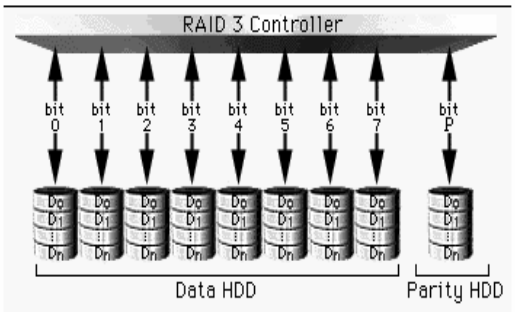
108

## RAID – Level 0




109

## RAID – Level Five




110

## Spinning Disks

- Jim Gray, "TerraScale Sneaker Net: Using Inexpensive Disks for Backup, Archiving and Data Exchange" (2002), Microsoft Technical Report 02-54

111

## Cost to Store One and Transfer One TB

**Table 3**: The relative cost of sneaker-net, using various media. The analysis assumes 6MBps tape, 10MBps CD/DVD and robots at each end to handle the media. Note that the price of media is less than the fixed robot cost.

| | Media | Robot$ | Media$ | TB read + write time | ship time | TotalTime /TB | Mbps | Cost (10 TB) | $/TB shipped |
|---|---|---|---|---|---|---|---|---|---|
| CD | 1500 | 2x800 | 240 | 60 hrs | 24 hrs | 6 days | 28 | $2,080 | $208 |
| DVD | 200 | 2x8000 | 400 | 60 hrs | 24 hrs | 6 days | 28 | $20,000 | $2,000 |
| Tape | 25 | 2x15,000 | 1000 | 92 hrs | 24 hrs | 5 days | 18 | $31,000 | $3,100 |
| DiskBrick | 7 | 1,000 | 1,400 | 19 hrs | 24 hrs | 2 days | 52 | $2,600 | $260 |

112

**Table 4:** The price list for a Terabyte Brick in a 3G host (GHz processor, GB of memory, and Gbps Ethernet)

http://pricewatch.com/

| Item | Price |
|---|---|
| Cabinet (Lian LiPC-68 USG 12 bay case) | 138 |
| Power Supply (EnermaxEG465AX-VD 431W) | 117 |
| Motherboard (Abit KX7A-RAID KT266A) | 108 |
| Cpu (AMD 2GHz Athlon XP 1800+) | 110 |
| 1 GB Memory (2x512MB PC2100 266MHz DDR) | 120 |
| 1 TB Disks (7xMaxtor EIDE 153GB ATA/133 5400RPM | 1,281 |
| Gbps Ethernet (SysKonnect SK-9D21 Gig copper) | 219 |
| DVD (Sony DDY1621 16x DVD) | 45 |
| Floppy & 3xIDE cables, Video Card | 57 |
| OS (WindowsXP Pro OEM) | 95 |
| Database (SQL Server 2000 MSDE) | 0 |
| Shipping | 50 |
| Labor | 100 |
| **Total** | **$2,440** |

Networking

113

## CERN Magnetic Media Cost Estimates

**Disk**

| components | total 3-year cost (CHF) | usable capacity (TB) |
|---|---|---|
| a Disk server - 20 IDE disks (each 72GB raw capacity) in a dual Pentium III server, 10 IDE channels, mirrored disk, GbitEnet network interface; assembled from standard components; price quoted includes installation at CERN, 3-year on-site warranty (disk replacement performed by CERN operations) | 17,000 | 0.72 |
| b Administration & operation (2K per year per server) | 6,000 | |
| c CERN engineering expertise - one engineer for 100 servers - (unit cost 1FTE=CHF150K) | 1,500 | |
| *Total 3-year cost* | *24,500* | |
| **Cost per year per TB** | **11,343** | |

**Tape**

| components | total 3-year cost (CHF) | usable capacity (TB) |
|---|---|---|
| a tape cartridge for STK 9940; assumed life 5 years | 150 | 0.07 |
| b additional StorageTek PowderHorn silo, for connection to an existing robotic system; 6K slots; the price is a price we have paid for a new add-on silo, but in practice we buy re-furbished silos for CHF 160K; warranty 1 year; lifetime conservatively assumed to be ten years. | 270,000 | 420 |
| c annual maintenance for one silo | 19,000 | |
| d annual operation costs - 1 technician for 4 silos at CHF 150K per FTE per year | 37,500 | |
| e annual cost of one silo - e=a/10+c+d | 83,500 | |
| f cost per year for slot - f=e/6000 | 14 | |
| Cost per year per TB | | |
| media | 429 | |
| robotics | 199 | |
| **Cost per year per TB** | **627** | |

114

## Optical Storage Media

- WORM optical disks
  - CDR
  - DVD
  - 5.25
- WORM optical tape

115

## Rewriteable Digital Storage Media

**Magnetic tape**

- Multiple tracks & read/write heads
  - High storage capacity
  - High data transfer rate
  - Non-sequential access
- Sony AIT Tape
- Storage TEK
- Super DLT
- LTO

**Phase Change Optical Media**

- Phase Change technology
- Plasmon 5.25 inch disk

116

## Current Magnetic Tape Technologies

- Super DLT
  - Current: 300 GB, 36 MB/Sec
  - Future: 1.75 TB, 125 MB/Sec
- SAIT
  - Current: 500 GB, 30 MB/Sec
  - Future: 1 TB, 60 MB/Sec
- LTO
  - Current: 400 GB, 80 MB/Sec
  - Future: 800 GB, 120 MB/Sec

117

## WORM Magnetic Media

**SEC 17a-4(f) Compliance**

Magnetic Disks

- EMC Centera
  - "Fixed Content"
  - API controls all writes
  - Document fingerprint
- NetApp LockVault
  - SnapLock
  - SnapVault

Magnetic Tape

- Sony
- IBM
- Storage Tek

118

## WORM Optical Media

**CD/DVD Form Factor**

- Removable but quasi direct access
- Storage (red laser)
  - DVD 4 – 4.7 GB
  - DVD 9 – 8.5 GB
  - DVD 10- 9.4 GB
  - DVD 18- 17 GB
- Transfer rate
  - 3 MB/sec
- Cost
  - $5/GB
- Storage (blue laser)
  - Five times greater

**5.25 inch WORM**

- Removable/library
- Storage (red laser)
  - IBM 3995 5.25 GB
- Transfer rate
  - 4.2 MB/second
- Plasmon UDO (blue laser
  - 30 to 60 GB storage
  - 8 to 18 MB per second
- Cost
  - $2.00/GB to.50/GB

119

## What Storage Media Should Be Used?

- Selection criteria
  - High storage capacity
  - High data transfer rate
  - Twenty years life expectancy
  - Established and stable market place presence
  - Affordability
  - Suitability

120

## English Heritage

- 360,000 digital photos of historic buildings and artifacts
- What storage media to use
  - CD
    - 660 MB
    - 18,563 CDs
    - Labels
  - DLT
    - 35 GB
    - 306 tapes

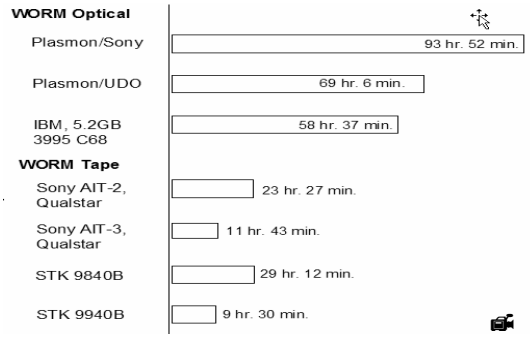
121

## Worm Optical v. Worm Magnetic

| Technology | Drive/Media | | Robotic Library | |
|---|---|---|---|---|
| **WORM Optical** | Model | Capacity (GB) | Model | Media Slots |
| Plasmon/Sony 5.25" MO CCW-WORM | Sony | 9.1 | G638 | 638 |
| Plasmon 5.25" UDO | Plasmon UDO | 30.0 | G638 | 638 |
| IBM 5.25" MO CCW-WORM | CCW-WORM | 5.2 | 3995 C68 | 258 |
| **WORM Tape** | | | | |
| Sony AIT-2 | Sony AIT-2 | 50.0 | Qualstar 412600 | 600 |
| Sony AIT-3 | Sony AIT-3 | 100.0 | Qualstar 412600 | 600 |
| STK 9840B | STK 9840B | 20.0 | L700/e | 678 |
| STK 9940B | STK 9940B | 200.0 | L700/e | 678 |

122

## WORM Optical v. WORM Magnetic




## Take Away

- Magnetic media more robust than optical
- Magnetic tape rather than “spinning disks”
- High data transfer rate

124

## File Formats

## What File Formats Do

- Tell operating system how to interpret 1s and 0s
- Specify the internal logical arrangement of digital objects
- Provide special instructions - compression algorithms
- Provide information understood by specific application software

126

## Types of File Formats

- Information content specific
- Proprietary v. non-proprietary

127

## Partial List of File Formats Currently in Use

- Text
- Vector Graphics
- Graphic Images
- Graphic Images (compression)
- Image File Header
- Audio

128

## Text File Formats

- Plain ASCII text
- Native word processing
- Rich Text Format (RTF)
- Standard Generalized Markup Language (SGML)
- Hypertext Markup Language (HTML)
- Extensible Markup Language (XML)
- Portable Document File Format (PDF-PDF/A)
- Open Document (XML)

129

## Vector Graphics File Formats

- Initial Graphics Exchange Specification (IGES)
- Computer Graphics Metafile (CGM)
- Scalable Vector Graphics (SVG)

130

## Graphic Image File Formats

- Tagged Image File Format (TIFF)
- Graphics Interchange Format (GIF)
- Portable Network Graphics (PNG)
- Portable Document Format (PDF)

131

## Graphic Image Compression Algorithms

- Lampel-Ziv-Welch (LZW)
- Group 4 (ITU)
- Joint Bi-Level Image Group (JBIG)
- Joint Photographic Experts Group (JPEG)
- Motion Picture Experts Group (MPEG)

132

## Audio File Format

- Motion Picture Experts Group
- MPEG - 3

133

## Considerations in Selecting File Formats

Making Sense of the Alphabet Soup

- Processibility considerations
- Technology considerations

134

**FILE FORMAT EVALUATION SUMMARY**

**TEXT**

| Format | Multi-media | Navigability | Confidentiality | Transferability | Integrity | Standard | Persistence | Storage |
|---|---|---|---|---|---|---|---|---|
| SGML | No | Yes | No | Yes | No | International | Yes | Nominal |
| XML | No | Yes | No | Yes | No | Internet | Yes | Nominal |
| HTML | No | Yes | No | Yes | No | Internet | Yes | Nominal |
| PDF | Yes | Yes | Yes | Yes | Yes | De Facto | Yes | Decrease |
| PDF/A | No | Yes | No | Yes | No | International | Yes | Nominal |
| RTF | No | No | No | Yes | No | De Facto | Potential | Increase |
| ASCII | No | No | No | Yes | No | International | Yes | Nominal |

135

**FILE FORMAT EVALUATION SUMMARY**

**VECTOR GRAPHICS**

| Format | Multi-media | Navigability | Confidentiality | Transferability | Integrity | Standard | Persistence | Storage |
|---|---|---|---|---|---|---|---|---|
| IGES | No | No | No | Yes | No | International | Yes | Nominal |
| CGM | No | No | No | Yes | No | International | Yes | Nominal |
| SVG | Yes | Yes | No | No | No | Internet | Potential | Nominal |

136

**FILE FORMAT EVALUATION SUMMARY**

**GRAPHIC IMAGE COMPRESSION**

| Format | Multi-media | Navigability | Confidentiality | Transferability | Integrity | Standard | Persistence | Storage |
|---|---|---|---|---|---|---|---|---|
| LZW | NA | No | No | Yes | No | De Facto | Uncertain | Decrease |
| Group 4 | NA | No | No | Yes | No | De Facto | Uncertain | Decrease |
| JBIG | NA | No | No | Yes | No | International | Potential | Decrease |
| JPEG | No | No | No | Yes | No | International | Yes | Decrease |
| MPEG | No | No | No | No | No | International | Yes | Decrease |

137

## XML Evaluation

- Strengths
  - Internet standard
  - Separates content from rendering
  - Navigable
  - Revisable
  - Transferable
  - Nominal file size increase
  - Domain specific extensions
- Strengths
  - Major market penetration
  - Widely used
- Weaknesses
  - No inherent protection of integrity
  - Requires semantics

138

## PDF Evaluation

- Strengths
  - Easy creation
  - Navigable
  - Backward compatible
  - Optional password protection
  - Authentication certificate
  - Published specification
  - Technology platform independent
  - Size decrease
- Weaknesses
  - Proprietary file format
  - Some transfer functionality

139

## The Preservation Problem

What is the best option for preserving electronic documents over archival time spans?

- TIFF?
  - Widely adopted
  - No access to underlying text without OCR
  - No mechanism for capturing logical structure
  - Difficult to create "born-digital" documents
- XML?
  - Good for describing logical structure, but not appearance
  - Many incompatible domain-specific schemas
- Native Format (e.g., MS Word)?
  - Several ubiquitous, but closed proprietary formats

140

## Desirable Properties of a Preservation Format

- Device independence
  - Can be reliably and consistently rendered without regard to the hardware/software platform
- Self-contained
  - Contains all resources necessary for rendering
- Self-documenting
  - Contains its own description
- Transparency
  - Amenable to direct analysis with basic tools

141

## Desirable Properties of a Preservation Format

Absence of technical protection mechanisms
- No encryption, passwords, etc.

Disclosure
- Authoritative specification publicly available

Adoption
- Widespread use may be the best deterrent against preservation risk

142

## PDF/A

- ISO 19005 (forthcoming)
- Subset of PDF for long-term preservation
- Nominal and full compliance
- Mandatory requirements

143

## The PDF/A Standard

- ISO 19005 specifies how to use the Portable Document Format (PDF) 1.4 for long-term preservation of electronic documents
- Applicable to documents containing character, raster, and vector data
- The standard does not address:
  - Processes for generating PDF/A files
  - Specific implementation details of rendering PDF/A files
  - Methods for storing PDF/A files

144

## ISO/TC 171/SC 2/WG 5

- ISO Joint Working Group (JWG) for PDF/A
  - ISO/TC 171/SC 2, *Document management applications – Application issues*
  - ISO/TC 130, *Graphic technology*
  - ISO/TC 46/SC 11, *Information and documentation – Archives/records management*
  - ISO/TC 42, *Photography*

145

## PDF/A Terminology

- PDF/A-1 refers to the format defined by Part 1 (ISO 19005-1) of the standard
- Part 2 (ISO 19005-2) will define PDF/A-2
- New Parts can be added to the PDF/A family of standards without obsoleting previous Parts

146

## PDF/A

- Non-proprietary standard
  - Based on a proprietary, but open format
- Developed by inclusive set of stakeholders
- Subject to rigorous technical review
- Minimal restrictions necessary to facilitate long-term preservation
- Not reliant on the existence of any particular reader

147

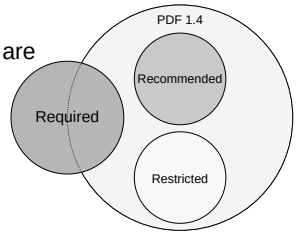
## PDF/A

- PDF/A is intended to address three primary issues:
  - Define a file format that preserves the static visual appearance of electronic documents over time
  - Provide a framework for recording metadata about electronic documents
  - Provide a framework for defining the logical structure and semantic properties of electronic documents

148

## PDF/A Requirements

- Conformance to PDF 1.4
- With features that are
  - Required
  - Recommended
  - Restricted




149

## PDF/A Conformance

- Two conformance levels
  - PDF/A-1a
    - Compliance with all requirements of 19005-1
    - Including those regarding structural and semantic tagging
  - PDF/A-1b
    - Compliance with all requirements of 19005-1 minimally necessary to preserve the visual appearance of a PDF/A file

150

## General

- Required
  - Conformance to 1.4 requirements
- Recommended
  - Linearization hints should be ignored
- Restricted
  - Document information dictionary must be consistent with XMP metadata
- Prohibited
  - Encryption
  - LZW compression
  - Embedded files
  - Optional content
  - Sound and movie media types

151

## Graphics

- Required
  - Device independent color
  - Embedded color spaces
- Restricted
  - Image dictionaries
  - **Separation** and **DeviceN** color spaces
  - Form XObjects
  - Extended graphics state
  - Rendering intents
- Prohibited
  - Reference XObjects
  - PostScript XObjects
  - Non-PDF 1.4 defined operators
  - Transparency

152

## Fonts

- Required
  - Fonts legally embeddable for unlimited, universal rendering
  - Embedded font programs
  - Embedded CMaps
  - Consistent font metrics
  - Unicode character map (For Level A conformance only)
- Recommended
  - Font subsets
- Restricted
  - Character encodings

153

## Metadata

Requires use of Extensible Metadata Platform (XMP)

- Proprietary, but open format
- Used for metadata creation, processing, and interchange
- Based on Resource Description Framework (RDF)
  - Open World Wide Web Consortium (W3C) standard
  - Cornerstone of Semantic Web
- Pre-defined schemas
  - Base, DC, DRM, DAM, Workflow, EXIF, PDF, PSD
- Defined extension mechanism
- Embedding rules
  - TIFF, JPEG, JPEG 2000, HTML, AI, PSD, PDF, …

154

## Metadata

- Required
  - Document level XMP metadata
  - Equivalent XMP metadata for all appropriate Document Information Dictionary properties
  - Embedded extension schema
  - Version and conformance self-identification
- Recommended
  - File identifier
  - File provenance
  - Font metadata
- Prohibited
  - XMP packet header **bytes** and **encoding** attributes

155

## Actions

- Required
  - Behavior for **NextPage**, **PrevPage**, **FirstPage**, and **LastPage** actions as defined in PDF 1.4
  - Reader mechanism to expose **GoToR** dictionary **F** and **D** keys, URI action dictionary **URI** key, and SubmitForm action dictionary **F** key
- Prohibited
  - **Launch**, **Sound**, **Movie**, **ResetForm**, **ImportData**, and **JavaScript** actions
  - Deprecated **set-state** and **no-op** actions
  - Named actions other than the 4 page navigation actions
  - Widget annotation or Field dictionary **AA** key

156

## Logical Structure (Level A Conformance Only)

- Required
  - Tagged PDF
  - Explicit word breaks
- Recommended
  - Tagging for pagination, layout, and page artifacts
  - "Strongly structured" block-level structural tagging
  - Natural language tagging
  - Alternative description, non-textual annotation, replacement text, and abbreviation/acronym expansion tagging

157

## PDF/A Summary

- ISO 19005-1 (should be available October 2005)
- File format standard
- One component of a comprehensive archival strategy
- Based on PDF 1.4
- Two conformance levels
  - Level A for structural/semantic tagging
  - Level B for appearance only

158

## PDF/A Summary Continued

- Emphasis on reliable and predictable rendering of static visual appearance
  - Do's: embed fonts, device-independent color, XMP metadata, tagging
  - Don'ts: encryption, LZW, embedded files, external content references, transparency, multi-media, JavaScript

159

## Open Formats

- OASIS
- XML v. 1.0
- XML Style sheet Language (XSL) v. 1.0
- XML Query Language (XQUERY) v. 1.0
- RTF
- Plain text
- PDF/A

160

## Take Away

- File format selection should be driven by recordkeeping requirements
- Avoid proprietary single vendor products
- Use main stream technology products
- Require transferability functionality
- XML, PDF, and PDF-A are good choices

161

## Metadata

## Metadata for Electronic Records

- Technical information
  - Creation and use
  - Software applications
  - File formats
- Business information
  - Business rules
  - Integrity rules
  - Access rights, authorization, etc
- Contextual information
  - Who, What, When, Why
  - Linkage between and among records
  - Preservation information
  - Audit trail

163

## Metadata – Crime Scene

- Data
  - Fingerprint matching suspect
  - DNA matching the victim
- Metadata
  - Who owned the knife? (interpretation of the fingerprints & blood)
  - When & Where was the knife found?
  - Who found it? Was there a chain of evidence?
  - Quality of the DNA match/fingerprint match
  - Quality of the lab performing the tests
- Metadata speaks to the trustworthiness of the data

164

## Metadata Approaches

- Dublin Core http://dublincore.org/
- PREMIS. Data Dictionary for Preservation Metadata http://www.oclc.org/research/projects/pmwg/
- METS. Metadata Encoding & Transmission Standard Library of Congress http://www.loc.gov/standards/mets/
- Getty Institute "Introduction to Metadata" http://www.getty.edu/research/conducting_research/standards/intrometadata/setting.html
- Minnesota Recordkeeping Metadata Standard http://www.mnhs.org/preserve/records/metadatastandard.html

165

## Preservation Metadata Model

- Purpose
  - Identify information that must be captured and preserved
  - Assure users that electronic records are trustworthy over time
  - Despite changes in technology
- Components
  - History profile
  - Reformat profile
  - Copy Profile
  - Conversion profile
  - Migration profile

166

## Preservation Metadata Model History Profile

- Establish basic information at transfer to repository
- Create an audit trail that documents processing actions
  - Who, What, When, How, Why

167

HISTORY PROFILE

(Repeatable)

| Scope | How | When |
|---|---|---|
| Name of file/record | Metadata/Data Entry | Processing |
| Date of creation | Metadata/Data Entry | Processing |
| Transfer Date | System | Processing |
| Physical Record Count | Metadata/System | Processing |
| Logical Record Count | Metadata/System | Processing |
| CRC (If Used) | Metadata/System | Processing |
| Hash Digest (If Used) | Metadata/System | Processing |
| Algorithm | Metadata/System | Processing |
| Software | Metadata/System | Processing |
| *Representation* | | |
| Text | Metadata/Data Entry | Processing |
| Image | Metadata/Data Entry | Processing |
| Vector | Metadata/Data Entry | Processing |
| Database | Metadata/Data Entry | Processing |
| GIS | Metadata/Data Entry | Processing |
| Audio | Metadata/Data Entry | Processing |
| Moving Image | Metadata/Data Entry | Processing |
| Web Page | Metadata/Data Entry | Processing |
| *Format* | | |
| ASCII | Metadata/Data Entry | Processing |
| UNICODE | Metadata/Data Entry | Processing |
| SGML | Metadata/Data Entry | Processing |
| XML | Metadata/Data Entry | Processing |
| HTML/XHTML | Metadata/Data Entry | Processing |
| SVG | Metadata/Data Entry | Processing |
| TIFF | Metadata/Data Entry | Processing |
| JPEG | Metadata/Data Entry | Processing |
| PNG | Metadata/Data Entry | Processing |
| PDF-PDF/A | Metadata/Data Entry | Processing |
| *Vendor Support for Format* | | Processing |
| Vendor | Metadata/Data Entry | Processing |
| *Access Dates* | | Processing |
| Date (repeatable) | Metadata/Data Entry | Processing |

168

## Section 6

## Operational Digital Archiving Programs

## Overview

- OCLC Digital Archives
- DSpace
- Washington Digital Archives

170

## OCLC Digital Archives

- Standards-based
  - OAIS
  - METS
- Content format
  - HTML
  - PDF
  - TXT
  - TIFF
  - JPEG
  - GIF
  - BMP

171

## Non-Custodial Digital Preservation

- Transfer
  - Web
  - Batch
- Service
  - Bit preservation
  - Full preservation

172

## OCLC Costs

Table 2. Annual Costs for Managed Storage at OCLC's Digital Archive, by Format

| | Format | Quantity | # GB | $ lowest unit cost | Total |
|---|---|---|---|---|---|
| Text | ASCII with encoding | 728,862 files | 2.09 | $15/GB* | $31 |
| | 1-bit page images | " | 70 | $15/GB* | $1,050 |
| | 8-bit page images | " | 3,161 | $15/GB | $47,415 |
| | 24-bit page images | " | 9,484 | $15/GB | $142,260 |
| Photos | 24-bit PhotoCD (~10.7 MB) | 3,000 images | 31.4 | $15/GB* | $ 471 |
| | 24-bit TIFF (~200 MB) | 1,200 images | 268 | $15/GB* | $ 4,020 |
| Audio | 96kHz/24 bit AIFF | 40 hours | 82 | TBD | -- |
| Moving images | "lossless" (62 Mbps) | 20 hours | 4,359 | TBD | -- |

* these $15/GB prices assume that owners have already reached the 1,000 GB threshold for their account; if these were first-time deposits, prices would increase to $32/GB for deposits of 101-1,000 GB and $60/GB for deposits totaling less than 100 GB.

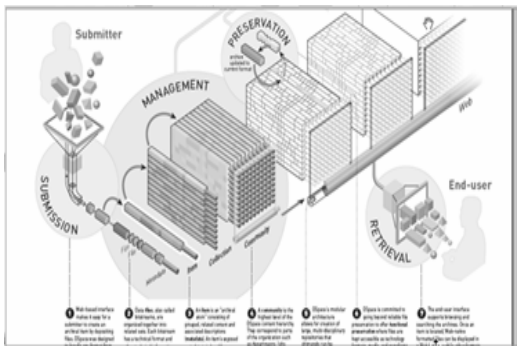
Stephen Chapman, "Counting the Costs of Digital Preservation is Repository Storage Affordable?"

173

## DSpace

- Digital repository system that captures, stores, indexes, preserves, and redistributes digital research material.
- MIT & HP
  - Open standards, Unix /Linux platform
  - No acquisition cost
- Functionality
  - OAIS compliant
  - Runs as is or modifiable to reflect institutional policy (e.g., formats)

174




175

## Digital Material Accepted

- Documents, such as articles, preprints, working papers, technical reports, conference papers
- Books
- Theses
- Data sets
- Computer programs
- Visualizations, simulations, and other models
- Multimedia publications
- Administrative records
- Published books
- Overlay journals
- Bibliographic datasets
- Images
- Audio files
- Video files
- Reformatted digital library collections
- Learning objects
- Web pages

176

## DSpace Functionality

- Ingests any digital object (in theory)
  - TIFF
  - PDF
  - XML
  - Text
- Exports
  - Images
  - Text

177

## DSpace Assessment

- Designed for academic library repositories
- Customization required for archives
- On-going archives resource requirements

178

## Washington Digital Archives

- Only operational digital state archives
  - Planning begun 1999
  - Opened in October 2004
- Legislature funding new records storage facility
  - SOS Office
  - Filing fees
- Initial cost of $14.8 million

179

## Washington Digital Archives (WDA)

- Well executed feasibility study
  - Recommended architecture, Windows based server
  - RAID-5
  - Near-line storage (HSM) later
  - Tape library
  - Convert electronic records to XML
    - Structured content
    - E-mail
  - Databases
    - Open Database Connectivity
    - XML version

180

## Legacy Records

- Support a conversion lab
  - Hardware and software for most common and still used formats
  - Convert to XML
- Outsource to conversion services
- Proof of concept testing
  - Stellent Outside In
  - XML Export Outside In
  - 97% converted with “equivalent look and feel”

181

## How WDA Will Work

- Ingested Web site of Governor Locke
- Three tier approach
  1. Agency high level technology capability
     - “content management system”
     - automatic transfer to WDA content management system
  2. Middle tier of technology capability
     - off the shelf applet to create metadata text
     - off the shelf applet to convert to XML
     - FTP to WDA server
     - Import into to WDA “content management database”
  3. Bottom tier of little technology capability
     1. Paper transfer of metadata
     2. Transfer metadata and records to a CD, tape, etc
     3. WDA staff will convert metadata and records to XML
     4. Import metadata and records into content management system

182

## WDA Observations

- Strong funding
- Helps to be in Washington
- Excellent feasibility study
- Realistic assessment of agencies and options but will they work?
- Do 3 technology levels map to record values?
- “Keeping bit stream alive”?
- Not easily transferable to other states

183

## Section 7

## Digital Archiving Programs To Watch

## Overview

- Georgia Digital Archives
- Smithsonian Institution/Rockefeller Archives Center
- Electronic Records Archives (NARA)

185

## Georgia Digital Archives (GDA)

- NHPRC funded demonstration project
- Collaboration
  - Pardon and Paroles Board (case files)
  - Georgia State Archives
  - $450,000
- Objectives
  - Implement records life cycle management of electronic records
  - Secure transfer of sensitive electronic records to GDA

186

## How Demonstration Works

- XML encoded clemency applications and supporting documentation
- Sensitive material that must be protected from public disclosure
- Capture of "archives ready" records
- PDF legacy case files
- FTP transfer
- Digital repository

187

## Infrastructure

- One temporary software engineer
- Existing technical and archives staff
- Use of TRIM for records capture
- Scalable low volume digital repository
- Transformation of Georgia State Archives

188

## Smithsonian Institution Archives & Rockefeller Archives Center

**Smithsonian Institution Archives**

- Quasi-federal environment
- Systems architecture in place
- Systems infrastructure in place
- Electronic records policy
- Technical expertise in place
- No resources to reach critical mass

189

## Smithsonian Institution Archives & Rockefeller Archives Center

**Rockefeller Archives Center**

- Documents important aspect of American society
- No regulatory authority or records policy
- Loose confederation of independent foundations
- All paper to no paper
- Wide spread use of databases, Email, and office productivity tools
- Virtually no technology infrastructure
- Outsourced IT
- One person IT shop
- Funds distribution
- Rockefeller University IT Department

190

## SIA/RAC Collaboration Project

- Objectives
  - Identify records of archival value
  - Implement digital archives repository with off the shelf technology
  - Develop a methodology that is generalizeable to other organizational settings

191

## SIA/RAC Project Team

- Darwin Stapleton (RAC) & Ricc Ferrante (SIA)
- Two electronic records archivists
- Software engineer
- Two consultants
- Partially funded
- Web site with progress reports

192

## ERA/NARA

- Scope
  - Preserve any kind of electronic records
  - No record "will be left behind"
  - No dependence on specific software or hardware
  - Hundreds of TBs and PBs
- Complex records
  - Privacy/FOIA
  - Multiple repositories with a single interface
  - Security classified records

193

## How ERA Will Work

- Complex architecture and infrastructure
- Heavy reliance on technology neutral standards
  - XML templates for archives ready records
  - PDF/A
- Migration strategy
- Retain original bit stream
- Automate entire process

194

## Success Factors

- Open systems and standards
- Modularization & iteration
- Ingest without human intervention
- Redesign of records management model – "archives ready" and map to ERA

195

## Lockeed Martin Award

- Announced September 8, 2005
- $308 million project over six years
  - $30 million year one
  - Additional funding will be required
- Archives of the future
- ERA Advisory Committee established
- Should state archives rely on ERA to solve electronic records preservation
  - Level of scalability?
  - Probably 8 to 10 years out at best

196

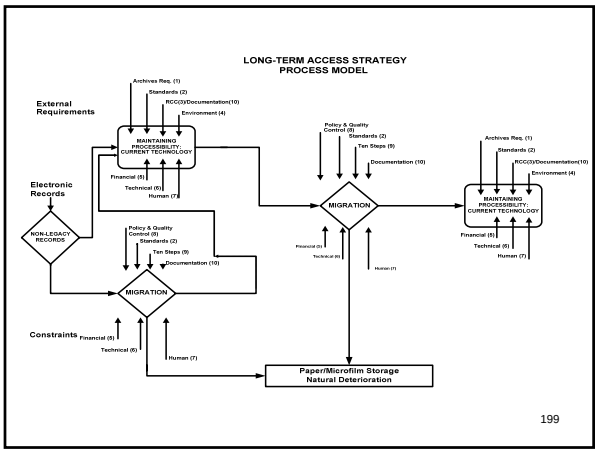
## Section 8

## A Digital Archiving Alternative

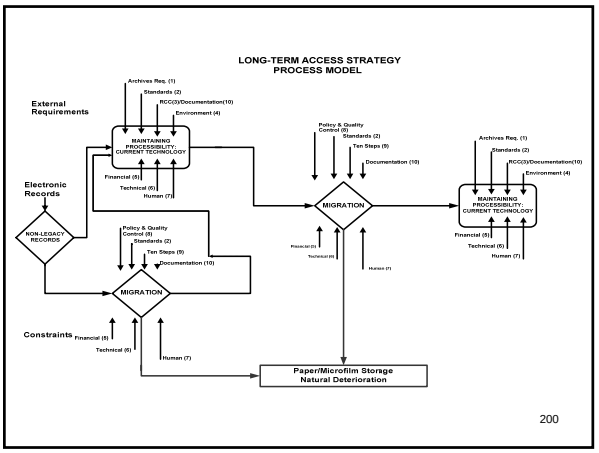
## Overview

- A digital microfilm and digital archiving strategies
- Two digital microfilm systems
- When to use digital microfilming

198




199




200

## Digital Microfilm

Advantages

- 500 years life expectancy
- Well understood technology
- Inherently technology neutral
- Immune to technology obsolescence
- Robust storage medium
- High quality images (reproducibility)
- High volume, low cost

Weaknesses

- Non-executable documents (non-processibility) render only
- Slow access time
- Does not support
  - databases
  - multimedia (animation)
  - hypertext
  - GIS
  - dynamic Web pages

201

## Digital Document Archive System

- Functionality
  - Converts digital documents to TIFF images
  - Writes TIFF images to 16 mm microfilm (7200/hour)
  - Index data & two level image marks
  - Retrieval and display
- Software
  - Archive Writer Interface Software (AWIS)
  - Lincoln EP/Fax PDF to TIFF plug-in
- Media
  - Reference Archives Media 1433/3433

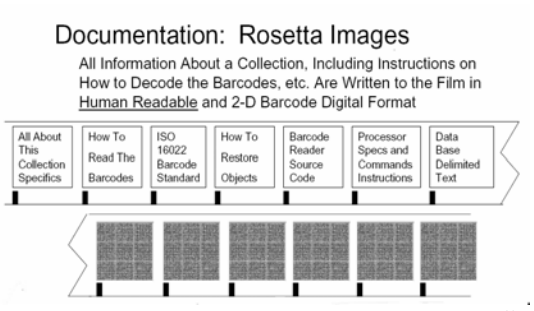
202

## Datasurance

- Non Proprietary Format: ISO Standard–2-D Barcodes are Simple & Standardized
- ISO non-proprietary format easily understood and readable without a computer peripheral
- Embedded Reed-Solomon Error Correction Code in the Barcode to Correct for Reading Errors and
- 2 Cyclical Redundancy Codes to Restore Accurately

203

## Datasurance (ACS)



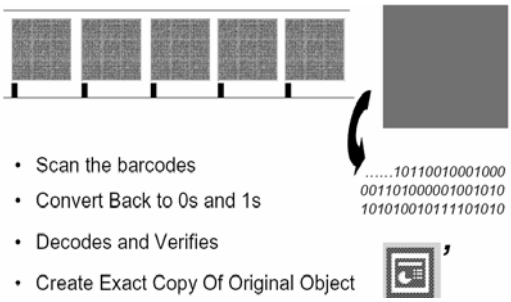

204

## How Datasurance Works

- Process Takes the Digital Object's 0s & 1s and Encodes Them Into a 2-D Barcode
- Process Assembles All The 2-D Barcodes of The Digital Object Into Binary Pictures and Writes Them to Film


205

## How Datasurance Works




- Scan the barcodes
- Convert Back to 0s and 1s
- Decodes and Verifies
- Create Exact Copy Of Original Object

206

## Datasurance

- Encoding, storage, decoding service
- Non-custodial preservation

207

## When To Consider Digital Microfilm

- Legacy records
- Page analog
- Long-term retention
- Robust finding aid
- Reproducibility satisfies
  - Regulatory compliance
  - Business needs
  - Historical accountability

208

## Section 9

## Putting Everything Together

## Practical Preservation Activities

## Multiple Levels of Digital Preservation

- "one size" does not fit all
- Categorize records
  - short-term v. long-term evidence
  - processibility v. reconstructibility
  - functions/activities that produce digital records
  - content is more important than "look and feel"

211

## Practical Preservation Activities

- Digital preservation mission statement
- Digital preservation policy
- Digital preservation strategy (ies)
- Digital preservation best practices

212

## Digital Preservation Mission Statement

- High-level statement
- Preserve and ensure access to usable and trustworthy records for as far into the future as possible
- The purpose of archives

213

## Digital Preservation Policy

- Describes at a high level how the mission statement will be carried out
- Specifies circumstances under which various preservation activities will be carried out
  - Custody – legal or physical
  - Secure repository (repositories)
  - Migration on demand, for example
  - Records integrity – quality control
  - Audit for compliance

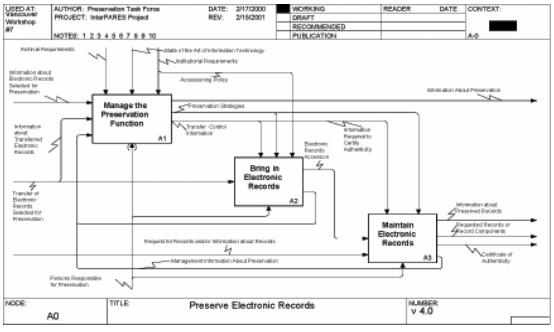
214

## Digital Preservation Strategy (ies)

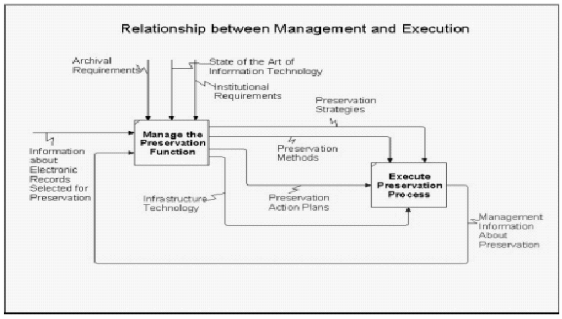
- Convergence of technology and policy
- OAIS and InterPARES preservation process models
- Data archaeology, migration, emulation
- Role of standards

215

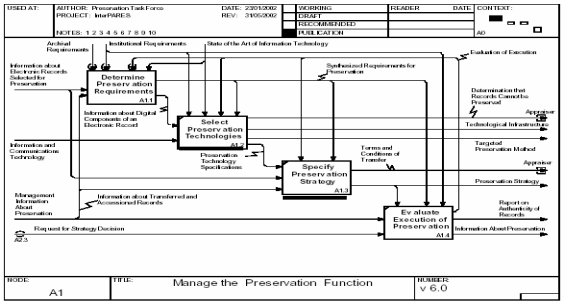
## InterPARES Electronic Records Preservation Model Overview



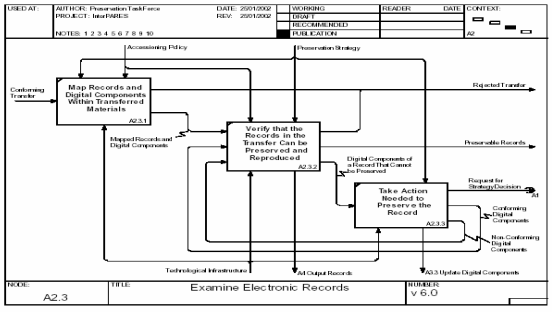
## InterPARES Electronic Records Preservation Model - 2



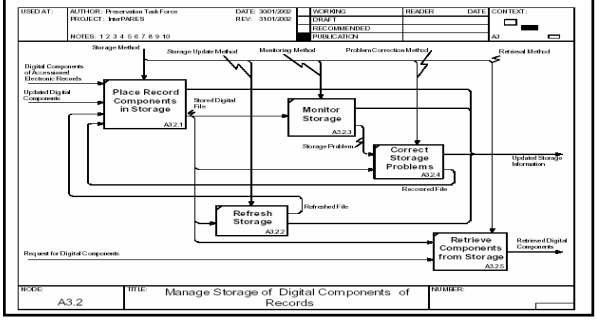

## InterPARES Electronic Records Preservation Model - 3




## InterPARES Electronic Records Preservation Model - 4




## InterPARES Electronic Records Preservation Model - 5




## Digital Preservation Best Practices

- Secure repositories in multiple geographic locations
- Media selection
- Media renewal
- Controlled storage environment
- Records integrity
- What not to do

221

## Multiple Secure Repositories

- Checklist for Certification of Trusted Digital Repositories
- Two or more geographically dispersed repositories
- Firewall
- Limited access to repository

222

## Keep Digital Records Readable

- Threshold issue
- Write properly up front
- Controlled environment in storage repository
- Periodic inspection of digital records
- Periodic transfer of digital records to new storage media

223

## Digital Storage Media

- Magnetic not optical
- Spinning disks
- Magnetic tape
- Preservation requirements

224

## Ensure Trustworthiness of Digital Records

- Maintain a secure repository
- Document digital preservation activities
- Use hash digest and digital time stamping as appropriate

225

## Implement "Best Practice"

- Research projects
  - San Diego Super Computer Center
  - NARA ERA
  - University of Michigan
  - Getty Institute
- Operational
  - Minnesota Historical Society
  - Washington Digital Archives
  - Georgia Digital Archives
  - Rockefeller Archives Center/Smithsonian Institution Archives

226

## What Not To Do

- Try to preserve everything
- Substitute quick fixes in lieu of long-term solutions
- Implement technology in the fringe of the market place

227

## Questions?

228

Missouri Secretary of State
Records Services Division

Records Management 573-751-3319
http://www.sos.mo.gov/records/recmgmt/

Local Records 573-751-9047
http://www.sos.mo.gov/archives/localrecs/

229

Charles M. Dollar
Thecdollar@cs.com
662-236-2479

230